Amser+: Accelerating Mobile Speech Emotion Recognition in IoT Environments with Mel Feature Compression

Yu Lu, *Student Member, IEEE,* Ran Wang, Dian Ding*, *Member, IEEE,* Yijie Li, *Member, IEEE,* Yongzhao Zhang, *Member, IEEE,* Lanqing Yang, *Member, IEEE,* Yi-Chao Chen, *Member, IEEE,* and Guangtao Xue, *Member, IEEE*

***Abstract*—Speech-based interaction systems are widely used in mobile devices like smartphones. With advances in deep neural networks, tasks such as speech emotion recognition (SER) enhance these systems' user-friendliness. However, deploying SER models on mobile devices is challenging due to their complexity and computational demands. While pruning can reduce complexity, it often compromises accuracy, and hardware accelerators like FPGAs are difficult to integrate into mobile devices. This paper proposes Amser+, a real-time speech emotion recognition framework using signal compression and task offloading. Amser+utilizes logarithmic Mel-filter bank coefficients (Fbank) and singular value decomposition (SVD) for feature extraction and compression. The compressed signal is only 6.25% of the original size, achieving 2.24x faster transfer rates and 55.35% energy savings compared to raw audio transmission. Despite the compression, the features preserve key audio information for text and emotion recognition, performed server-side. Experiments show a WER of 4.68% (Librispeech), 10.69% (CommonVoice), and 72.85% emotion recognition accuracy (IEMOCAP).**

***Index Terms*—Speech Emotion Recognition, Feature Compression**

I. INTRODUCTION

Speech is a prevalent interaction method in smartphones, stereos, and other IoT devices. The global speech and voice recognition market is expected to grow from $12.62 billion in 2023 to $59.62 billion by 2030 [1]. Unlike text, speech carries richer information such as emotion [2] and gender [3], [4]. Emotion recognition, in particular, enables intelligent systems to offer more personalized services [5]. For instance, in-car assistants monitor driver alertness and stress through speech, while smart customer service systems adjust their responses based on user emotion. In home healthcare settings, speech-based monitoring is increasingly used to detect mental health states or emotional fluctuations in elderly users. These scenarios require not only accurate emotion recognition but also low latency and high energy efficiency, given the constraints of mobile or embedded edge devices.

However, deploying real-time speech emotion recognition (SER) systems on such devices remains challenging. Deep neural networks [6]–[8], while accurate, are resource-intensive and typically require significant computational power, storage, and thermal headroom—resources that are often lacking in mobile environments. Moreover, applications such as voice assistants and in-car systems are highly sensitive to latency (e.g., responses must occur within 200–250 ms [9]–[11]), while also operating in noisy and bandwidth-constrained environments. Although researchers have reduced model complexity on mobile devices using techniques like branch pruning [12], weight sharing [13], tensor quantization [14], and knowledge distillation [15]–[17], these often reduce accuracy. Hardware solutions like GPUs [18], FPGAs [19], and ASICs [20], [21] improve computational capacity but are difficult to deploy on mobile devices due to size and power constraints.

We propose Amser+, a distributed speech emotion recognition framework using signal compression. Rather than compressing raw audio directly, Amser+ shifts the compression focus to the Mel-spectrogram domain, which serves as the primary feature in most downstream speech tasks. On mobile devices, the system computes Mel-filter bank (Fbank) coefficients and applies singular value decomposition (SVD) [22] to extract compact, low-rank representations. This strategy reduces the feature size to only 6.25% of the original audio, significantly lowering transmission and storage demands while preserving perceptually relevant emotional cues. Deploying real-time speech applications on mobile devices faces several challenges. First, mobile devices have limited computing power, making it hard to support complex neural networks. Second, IoT devices like smart speakers lack storage for long-term audio data and large models. Lastly, current emotion recognition models rely solely on dataset knowledge, limiting their accuracy.

We propose Amser+ to address these challenges by creating a real-time speech emotion recognition framework for mobile devices and servers. The system offloads deep neural network tasks to servers, reducing the computational and storage burden on mobile devices. It also compresses speech signals using Fbank features and SVD, minimizing storage needs. Contemporary methods [23]–[28] commonly use neural networks for emotion recognition, feature extraction, and data classification. Building on these approaches, we propose a novel multimodal model for emotion recognition. We first apply Automatic Speech Recognition (ASR) [6] to convert audio signals into text. We incorporate external knowledge using a pre-trained RoBERTa model to enhance emotion recognition accuracy further. Additionally, we employ text embeddings for extracting emotion-related features from the

* Corresponding author.

Y. Lu, R. Wang, D. Ding, L. Yang, Y. Chen, and G. Xue are with the Department of Computer Science and Engineering, School of Electronic Information and Electrical Engineering, Shanghai Jiao Tong University, Shanghai 200240, China. E-mail:{yulu01, wang_r, dingdian94, yanglanqing, yichao, gt_xue}@sjtu.edu.cn.

Y. Li is with the School of Computing, National University of Singapore, 117417, Singapore. E-mail:yijieli@nus.edu.sg.

Y. Zhang is with the Department of Computer Science and Engineering, University of Electronic Science and Technology of China, Chengdu 611731, China. E-mail:zhangyongzhao@uestc.edu.cn.

Chinese text and use a co-attention mechanism to fuse multimodal features effectively. Given the imbalanced distribution of emotion data in current datasets, such as IEMOCAP [29], and the fact that a single audio sentence may contain rich emotional expressions that a single emotion label cannot fully capture, we address these challenges by employing a MoCo-based [30] contrastive learning approach to train our model. This method helps improve the model's performance by better capturing the nuanced emotional information in the data.

Extensive experiments demonstrate the feasibility of deploying a real-time speech emotion recognition system on mobile devices. The key contributions of this paper are as follows:

- We propose Amser+, a speech emotion recognition system for edge mobile devices. Unlike traditional systems that offload all computations to the server, Amser+reduces transmission latency and optimizes resource usage on edge devices.
- We propose a feature extraction and compression module for audio signals, optimized for mobile devices. Using Fbank, the audio is converted into an acoustic spectrogram, with SVD applied to compress and filter out high-frequency redundant information.
- We constructed a multimodal neural network for speech emotion recognition based on the whisper [6], RoBERTa [7] models and MoCo-based training strategy.
- Extensive experiments show that compared to direct raw audio transfer, Amser+improves transfer rates by 2.24x, reduces energy consumption by 55.35%, and achieves a 6.25% file compression ratio. On the IEMOCAP dataset, it achieves 72.85% accuracy and an F1-score of 0.713.

II. RELATED WORK

A. Real-Time Mobile Computing Applications

Real-time mobile computing has become increasingly essential for latency-sensitive applications such as on-device speech processing [31]–[33], affective computing [34]–[36], and human–machine interaction [37]–[42]. Typical examples include voice-controlled smart home systems [43], where user commands must be processed with minimal latency; In-vehicle driver monitoring [44], [45], which requires real-time analysis of driver behavior to ensure safety; industrial equipment monitoring [46], where fault detection and predictive maintenance rely on rapid local computation; mobile health systems [47], [48] that continuously track physiological signals for early warnings; and augmented reality (AR) [49], [50] applications that demand instant sensor fusion and feedback. These scenarios [51]–[53] illustrate the growing need for mobile systems that can operate under strict latency, energy, and connectivity constraints while ensuring reliable, real-time performance.

B. Deep Neural Network Deployment

Deploying DNN models on edge devices is a common challenge in AI fields like NLP and computer vision. Solutions such as Vigil [54], Reducto [55], Filter-Forward [56], and Glimpse [57] implement selective data offloading to minimize latency based on feature type, filtering thresholds, and content. Cracking open the DNN [58] enhances video analytics through joint camera-cloud inference and continuous online learning. Elf [9] improves mobile deep vision by distributing inference tasks to multiple servers. Remix [59] optimizes object detection on edge devices with image partitioning strategies under latency constraints. Amser+offers a real-time speech emotion recognition framework via compression and task offloading.

C. Speech Emotion Recognition

Speech emotion recognition has been studied for multiple decades within both the machine learning and speech communities. In alignment with the prevailing research approach, scholars extract feature insights from audio data and subsequently employ these insights across a range of classifiers, including: hidden Markov models [60], convolutional recurrent network [61], SVM [62], hierarchical binary decision tree [63], gaussian mixture [64], nerual network [65]. Much of the aforementioned works relied on context to furnish additional information for correcting and inferring emotional content extracted from the data. The mining and analysis of emotional information from single-sentence audio data can pose more significant challenges. Xu et al. [26] introduced an attention-based network designed for aligning textual and audio information, along with feature extraction. Yoon [27], [66] presented a groundbreaking deep dual recurrent encoder model that seamlessly merges text data and audio signals. This model employs a pair of recurrent neural networks (RNNs) to holistically encode the information. Delbrouck [67] et al. proposed a transformer-based joint-encoding model called UMNOS for single-sentence emotion recognition and sentiment analysis.

III. PRELIMINARY STUDY

In speech recognition tasks, methods like MFCC or Fbank are commonly used to extract two-dimensional features from audio signals through windowed sampling. For example, OpenAI's Whisper [6] uses Fbank to extract acoustic spectrograms from audio, followed by a transformer-based encoder-decoder model to convert the spectrogram into text labels.

Features extracted through Fbank often contain redundant information, with high-frequency details offering limited utility in systems like Whisper. Similar to image compression, where high-frequency details can be removed without losing key information, we propose using the SVD algorithm to compress acoustic spectrograms. This preserves low-frequency features while reducing dimensionality for better identification and classification.

We verify the efficacy of SVD for compressing audio features within the Whisper speech recognition framework. In the Whisper framework, the speech signal $s \in \mathcal{R}^t$ undergoes extraction by Fbank to yield the acoustic spectrogram feature matrix $f \in \mathcal{R}^{m \times n}$:

$$f = Func_{Fbank}(s) \tag{1}$$

Let $k = min(m, n)$, then we compute the SVD of matrix f:

$$\begin{aligned} &f = Udiag(S)V^H \\ &U \in \mathcal{R}^{m \times k}, S \in \mathcal{R}^k, V \in \mathcal{R}^{n \times k} \end{aligned} \tag{2}$$

where $diag(S) \in \mathcal{R}^{k\times k}$, V^H is the conjugate transpose when V is complex, and the transpose when V is real-valued, and the matrices U, V are orthogonal in the real case, and unitary in the complex case. In this scenario, singular values S are sorted in descending order and are distinct. Denoting them as $\sigma_1 > \sigma_2 > \sigma_3 \cdots > \sigma_k$. Then f can be expressed as the following decomposition:

$$f = Udiag(S)V^H = \sum_{i=1}^{k} \sigma_i \begin{pmatrix} | \\ u_i \\ | \end{pmatrix} \begin{pmatrix} - & v_i & - \end{pmatrix} \quad (3)$$

where $U = (u_1, u_2, \ldots, u_k)$ and $V^H = \begin{pmatrix} v_1 \\ v_2 \\ \vdots \\ v_k \end{pmatrix}$.

Considering that the contribution of these singular values to the matrix shrinks sequentially, then according to the Eckhart-Young theorem [68],we can take the compression approximation of the acoustic spectrogram features:

$$f \approx f^{'} = \sum_{i=1}^{r} \sigma_i \begin{pmatrix} | \\ u_i \\ | \end{pmatrix} \begin{pmatrix} - & v_i & - \end{pmatrix} \quad (4)$$

where $r \in \mathcal{N} \cap [1, k]$, and $\frac{r}{k} \in [\frac{1}{k}, 1]$ denotes the compression rate for acoustic spectrogram features. In contrast to the original method where we needed to store U, S, V to recover f, now we only need to save $U^{'} \in \mathcal{R}^{m\times r}, S^{'} \in \mathcal{R}^{r}, V^{'} \in \mathcal{R}^{r\times n}$ to recover $f^{'}$, resulting in a saved matrix size equal to $\frac{r}{k}$ of the original.

Subsequently, we compress the Librispeech [69] and CommonVoice [70] datasets at various compression rates and assess the Whisper system's performance in recognizing the compressed acoustic spectrogram features. The Librispeech dataset is a large-scale corpus of read English speech, widely used for evaluating automatic speech recognition (ASR) systems. It contains approximately 1,000 hours of transcribed speech from audiobooks and is designed to evaluate systems in terms of both word error rate (WER) and transcription quality. The CommonVoice dataset, created by Mozilla, is an open-source initiative aimed at collecting a wide variety of speech samples from diverse speakers. It contains over 7,000 hours of audio data in multiple languages and serves as a benchmark for testing ASR systems across various domains and accents. As a common metric of the performance of a speech recognition or machine translation system, word error rate (WER) is employed to evaluate the performance of whisper on both datasets and can be caculated by the following formulation:

$$WER = \frac{S + D + I}{S + D + C} \quad (5)$$

where S is the number of substitutions, D is the number of deletions, I is the number of insertions and C is the number of correct words. The results depicted in the Fig. 1 demonstrate that when the compression rate exceeds 10%, the Whisper system exhibits commendable speech recognition performance even for compressed speech.

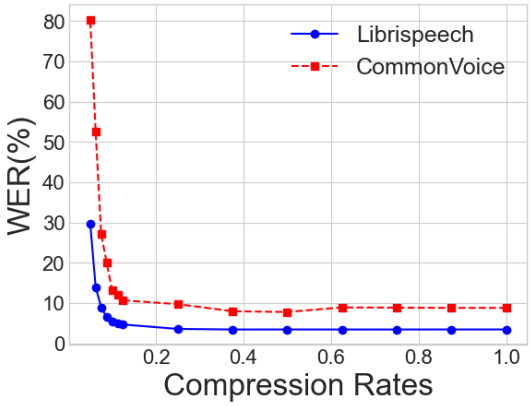

Fig. 1. Impact of Compression Rate for Whisper.

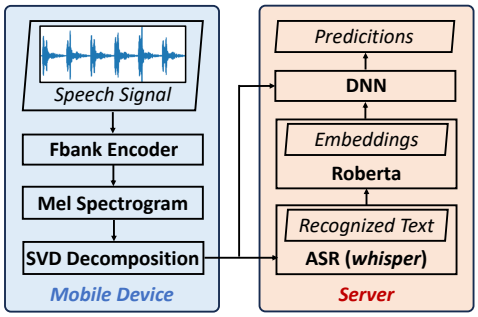

Fig. 2. The system architecture of *Amser+*.

Although edge devices may lack the computational power for large-scale models, extracting Fbank features and compressing them for server transmission is feasible. Compared to direct audio file transmission, sending compressed spectrograms reduces bandwidth usage and communication time. Previous studies show that SVD-based compression at 12.5% for spectrograms (6.25% for audio files) minimally impacts ASR performance. Amser+will further verify that this compression rate maintains accuracy in speech sentiment analysis.

IV. SYSTEM

We present Amser+, a real-time speech emotion recognition system. It consists of two parts: the mobile device acquires speech and extracts features using a Fbank encoder to output a Mel Spectrogram, which is then compressed to reduce storage. The compressed features retain text and emotion information, and the server performs text and emotion recognition using Whisper and multimodal networks. The system architecture is shown in Fig. 2.

A. Signal Preprocess

1) Feature Extraction: The mobile device extracts Fbank features (filter bank features) from the user's speech using a series of preprocessing steps designed to capture the essential characteristics of the audio signal. The process begins with pre-emphasis, which amplifies the higher-frequency components of the signal, helping to balance the energy across

the entire frequency range and reduce the effects of low-frequency noise. This step is crucial for enhancing the clarity of high-frequency signals, which are often important for speech recognition tasks. Next, the signal is split into frames with overlapping segments, which ensures that no abrupt changes occur between consecutive frames. The frame overlap effectively captures the temporal continuity of the speech, reducing the risk of losing important transitions in the audio signal. This step is particularly important for maintaining smooth transitions in speech analysis. Each frame is then windowed using a Hamming window to minimize edge effects and reduce spectral leakage. The Hamming window smooths the signal, ensuring that the transition at the boundaries of each frame does not introduce artifacts that could negatively impact feature extraction. The next step is the Short-Time Fourier Transform (STFT), which converts the time-domain signal into the frequency domain. The STFT breaks the signal into smaller segments, allowing the model to analyze frequency components over time. This transformation enables the model to capture both the spectral content and temporal evolution of the speech signal. Finally, the signal is passed through Mel filtering to convert the frequency-domain representation into a scale that more closely mimics human auditory perception. The Mel scale compresses high frequencies while preserving the perceptually significant features of the signal, ensuring that the extracted features align with the way humans perceive sound. This makes the features more suitable for emotion recognition, as it emphasizes the frequencies most relevant for distinguishing emotional cues in speech.

2) Signal Compression: In addition to the challenges posed by limited computational power, storage space is another critical constraint for mobile devices that cannot be overlooked. Given the need to handle large volumes of speech data in real-time, efficient storage management becomes crucial for the performance of the system. To address this, the system employs Singular Value Decomposition (SVD), as described in detail in Sec III, to effectively compress the speech features. SVD is used to reduce the dimensionality of the extracted features, discarding less important components while retaining the most significant information necessary for emotion recognition. This compression process not only minimizes storage requirements, but also helps with de-noising, removing irrelevant or noisy data that may interfere with accurate emotion classification. By focusing on preserving the key elements of the speech features—such as textual information derived from Automatic Speech Recognition (ASR) and the emotional cues embedded in the speech signal—the system ensures that only the most relevant data is retained. This balance between compression and feature preservation ensures that the mobile device can operate efficiently while still maintaining high accuracy in recognizing emotions. Ultimately, this approach enables the system to handle large amounts of data in real-time, while effectively managing the trade-off between storage limitations and the need for rich, high-quality features for emotion recognition.

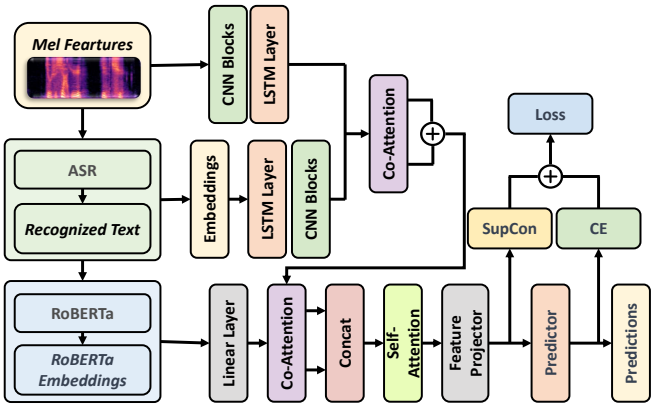

Fig. 3. The multi-modal model for emotion recognition.

B. Emotion Recognition

Here, we describe our emotion recognition model. This model employs three distinct modalities of data as input sources: Mel, word embeddings, and RoBERTa-encoded embeddings. Initially, each modality is processed separately. Subsequently, all the features from the various input modalities are combined using a *co-attention layer.* Finally, *Linear layers* are employed to produce the predictions. The overall model structure is shown in Fig. 3.

1) Modality input: First, the compressed features, derived from the Mel spectrogram (via Short-Time Fourier Transform, or STFT), effectively capture the temporal dynamics of signal energy changes, aligning with human auditory perception. These features provide a compact yet informative representation of the speech signal, preserving critical frequency components that are essential for emotion recognition.

After SVD decomposition on the mobile edge device, the Mel features are reconstructed on the server, retaining the key characteristics of the audio signal along with the semantic information required for emotion classification. This approach ensures that only the most relevant features are transferred, reducing the amount of data while maintaining the integrity of the emotional cues essential for accurate recognition.

To further enrich the representation, we leverage the Whisper [6] model for Automatic Speech Recognition (ASR) to convert the speech signal into its textual form. The transcriptions generated by ASR serve as an additional source of information, complementing the audio features. Next, we utilize text embeddings to directly learn the semantic features from the text input. Specifically, we employ a 300-dimensional pre-trained GloVe embedding [71] obtained through spaCy. This embedding encodes the transcription into fixed-length vectors, providing a dense representation of the semantic meaning of the spoken content.

In parallel, we integrate a pre-trained RoBERTa model to extract higher-level transcription features, allowing the model to incorporate external knowledge from large corpora. This enables the model to understand contextual nuances and semantic relationships within the text, which can significantly enhance emotion recognition, especially for more complex or

ambiguous emotional expressions.

By combining these multimodal feature extraction techniques—Mel features, text embedding, and external knowledge integration—we create a rich, multi-faceted representation of the speech signal, improving the model's ability to accurately recognize emotions.

2) Modality pre-process: After retrieving the mel-spectrogram of the audio signals, we apply a classic Conv-BatchNorm-ReLU structure to extract features in both the time and frequency dimensions. Then, an LSTM layer is applied to extract deeper features in the time dimension. Additionally, the word embeddings have a better time structure and are more straightforward in each time slot. Hence, an LSTM is applied to the word embeddings before using a 1D-convolution layer to incorporate the information from the entire timeline. The feature extracted from BERT is a 768-dimensional vector. As it is already well-structured and contains abundant information, we applied a Linear layer to modify its size for subsequent multi-modal fusion and information compression.

3) Multi-modal fusion: Given the presence of three modalities, we need two rounds of fusion to combine all the information extracted from these different modalities comprehensively, and determining the order of fusion is a significant consideration. In our model, we first fuse the audio features and word embedding features. Their akin temporal structures make them suitable for initial fusion, as this process enhances the temporal dimension by leveraging their shared characteristics to amplify common information and compensate for missing data unique to one modality. Subsequently, the time-structured feature mentioned earlier is fused with the BERT-encoded feature, incorporating external knowledge from the outside world to in-dataset knowledge. In each fusion, there are two stages: extracting additional features from one modality with knowledge from another modality and then merging these additionally extracted features into a single representation.

In the first stage, we employed the co-attention layer to convey the presence of another modality to each modality. The structure of *co-attention layer* is as shown in Fig. 4. Inspired by [72], we employed the Encoder-Decoder structure to stack multiple layers of attention modules. In the co-attention layer, the first modality employs self-attention alone to extract deeper information from itself. Following that, the second modality goes through a self-attention operation, during which a guided-attention step is conducted to extract more information while considering both modalities. In contrast to simply using the output of the self-attention from another modality at the same depth as the input for guided-attention, leveraging the final output of the Self-attention layers can offer more enriched information and a more accurate guide. Both self-attention and guided-attention are based on the attention mechanism [73]. The attention module aids in constructing a holistic perspective of the entire time span during the speech. The attention consists of a query q, a key k and a value v:

$$Attention(q, k, v) = softmax(\frac{qk^T}{\sqrt{k}})v \tag{6}$$

In the self-attention, all of q, k and v are from the same modality. However, in guided-attention, the v and k are from the same modality while q is from another different modality.

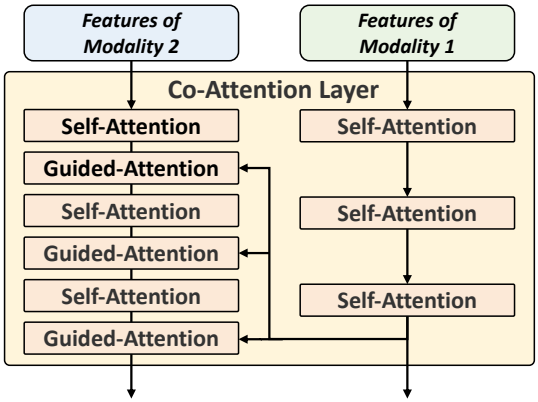

Fig. 4. The architecture of the co-attention layer.

The first stage of the two fusion is the same, yet they diverge in the second stage. Considering the similarity of time structures, for the fusion between features from audio data and word embeddings, we employ a straightforward element-wise addition. This approach enhances their temporal structure and reduces the feature size compared to concatenation. In the second fusion, the features are dissimilar and lack a shared temporal structure, which leads to lossy and disorganized information when using element-wise addition. Consequently, concatenation is employed to retain more information, which is crucial for effectively leveraging both in-dataset knowledge and external-world knowledge. Following the ultimate fusion, we applied additional self-attention to comprehensively process the collective information from all modalities and proceed to make predictions using a two-layer MLP.

4) Contrastive learning: Through our examination of misclassified cases in current state-of-the-art models, we identified that the ambiguity in the emotions expressed by actors is another factor hindering the model from learning accurate features. It is common to observe that a person's emotions can be complex, even involving contradictory feelings simultaneously. However, datasets with labels assigned to a single emotion as the ground truth may be misleading in capturing the presence of other coexisting emotions. Furthermore, employing traditional cross-entropy loss during model training mechanically steers the model to predict a probability of 1 only for the labeled emotion, penalizing predictions with non-zero probabilities for other emotions. This situation can significantly perplex the model, especially in cases where multiple emotions coexist. Moreover, stemming from naturalistic conversations in daily life, our dataset exhibits an imbalanced distribution of labels. Specifically, there is a pronounced prevalence of sentences labeled as neutral, contrasting with a scarcity of instances labeled as surprise.

Consequently, we advocate for the implementation of a contrastive learning loss as a regulatory measure to alleviate the impact of multiple emotions and mitigate data imbalances. Contrastive learning is a training technique that originated from unsupervised learning. Supervised learning studies [74]

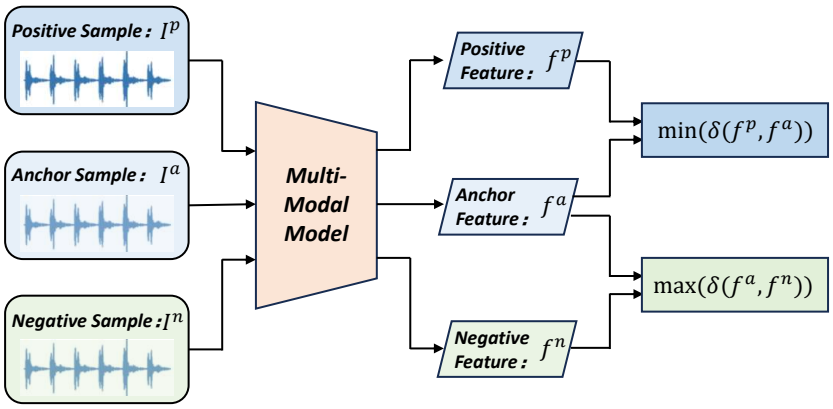

Fig. 5. The pipeline of the contrastive learning.

have also demonstrated its effectiveness, utilizing samples from the same class as positive samples and others as negative samples. The loss used in [74] is following:

$$L_{SupCon} = -\sum_{i \in I} \frac{1}{|P(i)|} \sum_{p \in P(i)} \log \frac{exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_p / \tau)}{\sum_{\alpha \in A(i)} exp(\boldsymbol{z}_i \cdot \boldsymbol{z}_\alpha / \tau)} \tag{7}$$

Here, I is the set of classes, $A(i)$ is the batch of samples contrasting with feature $\boldsymbol{z}_i$, $P(i)$ is the set of positive samples of feature $\boldsymbol{z}_i$ in $A(i)$, i.e. samples with the same label.

The loss function is characterized by a vague description, suggesting that the feature extracted from a given sample should exhibit proximity to features extracted from positive samples while maintaining distance from features of other negative samples. Unlike traditional supervised learning, which prescribes a specific point in a lower dimension for a sample, contrastive learning defines positions in high-dimensional space that a sample should either approach or diverge from. This can mitigate the impact of labels, thereby diminishing the influence of multiple emotions.

As depicted in Fig. 5 and Fig. 3, the contrastive learning loss is computed from the feature projector's output, whereas the conventional cross-entropy loss relies on the output of the predictor. The feature projector and the predictor are both a one-layer MLP. Therefore, the final loss can be represented as

$$L = \frac{L_{CE} + \alpha \cdot L_{SupCon}}{1 + \alpha} \tag{8}$$

where α is a hyperparameter to control the importance of contrastive learning loss in the final loss.

5) Data Augmentation.: We formulate data augmentation strategies to mitigate the impact of noise, thereby improving the overall generalization of the model. In detail, we augment the audio signals in three ways: adding noise based on SNR, applying pitch shifts, and employing time stretching. When adding noise to the audio feature, we use an SNR of $30dB$, and randomly initialize the noise in Gaussian distribution. The pitch shift and time stretch are implemented by the *librosa*. In IEMOCAP, to increase the contrastive samples, we take advantage of the Dropout layers in our model. We run the prediction twice in one epoch to generate different features from the same sample. Also, as described in the previous section, we adopted MoCo [30] with size 16384.

V. EVALUATION

A. Dataset

We use the IEMOCAP (Interactive Emotional Dyadic Motion Capture) dataset [29] and VCEMO [75], collected by the University of Southern California (USC) and Shanghai Voicecomm Information Technology Co., respectively, to evaluate our Amser+ system and train and test our model.

IEMOCAP: The dataset consists of 5 sessions featuring 10 actors (5 male, 5 female), with approximately 12 hours of total recording time. These interactions are scripted to evoke a range of emotions, and the recordings include both spontaneous and acted emotional expressions. The emotion categories in IEMOCAP are anger, happiness, sadness, fear, disgust, neutral, and a few mixed emotions. The audio recordings are annotated with emotion labels at the sentence level, making them ideal for studying speech-based emotion recognition. The dataset contains both acted and natural emotional speech, providing a rich resource for training and evaluating emotion recognition models. Additionally, it includes transcriptions of the spoken content, which makes it well-suited for multimodal approaches that integrate both speech and text analysis. IEMOCAP has been widely used for training, validating, and testing emotion recognition models, establishing it as a reliable resource for exploring techniques in speech emotion recognition (SER) and multimodal emotion recognition.

VCEMO: VCEMO is a recently proposed multi-modal emotion recognition dataset tailored for Chinese voiceprint-based applications. It comprises 7,477 single-sentence utterances collected from over 100 native speakers across diverse dialects and spontaneous conversational settings, thereby offering a rich variety of emotional expressions and acoustic features. Unlike prior datasets that rely on professional actors,

VCEMO reflects real-world dialogue scenarios, with each utterance annotated by experts into six emotion categories: angry, fear, happy, neutral, sad, and surprise. VCEMO is particularly valuable for developing and benchmarking robust, context-aware emotion recognition systems in Mandarin Chinese.

B. Experimental Setup

1) Device: **Sever.** We utilize a server equipped with 188 GB of RAM and a $48.0GB$ VRAM's NVIDIA A40 as our evaluation system for model training and testing.

Client. Redmi Note 12 Pro equipped with 8 GB of RAM and Mediatek dimensity 1080 is used as a system client for audio file processing and compression.

2) Model training: The model was trained for 100 steps with a batch size of 256 to ensure efficient data processing and stable gradient updates. The Adam optimizer [76] was used for its adaptive learning rate mechanism, which helps improve convergence efficiency. The learning rate was set to 1e-5, a typical choice for fine-tuning speech recognition models, enabling effective parameter updates without overshooting. The weight decay was set to 0, as no additional regularization was needed beyond other training strategies to avoid overfitting.

These hyperparameters were chosen based on standard practices in speech-related tasks, balancing stability and convergence speed. The smaller learning rate, paired with a relatively large batch size, facilitates gradual convergence, especially in the context of emotion recognition in speech data.

For the contrastive learning component, the temperature parameter t of the contrastive loss function was set to 1. This value was chosen to strike an optimal balance between the separation of positive and negative pairs in the embedding space, allowing the model to effectively learn both the structure of the data and the subtle emotional cues in the speech signals.

The training process was conducted using PyTorch [77] on an NVIDIA A40 GPU. This hardware choice facilitated efficient parallel computation, enabling faster training while maintaining high performance during multiple training steps. During the training process, key metrics such as accuracy, loss, and emotion recognition performance were continuously monitored to ensure consistent model improvement.

C. Baseline Method

To evaluate the performance of our proposed system, we compare it against four baseline methods, each representing different approaches for speech signal compression and transmission. These baselines are designed to showcase the effects of various compression techniques on emotion recognition performance.

- Raw Audio (WAV) Transmission: The first baseline represents a system that directly transmits the raw audio file, typically in WAV format, without any compression. While this method ensures high-quality audio transmission, it comes with the drawback of large file sizes, which increases both storage and bandwidth requirements. Despite the lack of compression, which preserves the original signal quality, it may not be the most efficient for real-time applications, especially on resource-constrained mobile devices.
- MP3/AAC Compression: The second baseline utilizes lossy audio compression techniques, such as MP3 or AAC, to reduce the size of the audio file. These formats achieve significant compression by removing less perceptible audio components, balancing between compression ratio and audio quality. While MP3 and AAC compression help reduce file size and bandwidth usage, they can also lead to some loss of information, potentially affecting the quality of the extracted features for emotion recognition, especially for subtle emotional cues.
- Mel Spectrogram Compression using Interpolation: The third baseline involves compressing Mel spectrogram features using interpolation techniques. In this method, the Mel spectrogram is downsampled using interpolation algorithms to reduce its size. Interpolation maintains a close representation of the original Mel spectrogram while reducing its dimensionality. However, this approach may not preserve the finer details of the spectrogram, which could impact the ability of emotion recognition systems to identify subtle emotional variations in the speech.

1) Evaluation Metrics: To assess the performance and effectiveness of our model for speech emotion recognition, we use several evaluation metrics. These metrics help us to capture not only the accuracy of the model but also its operational efficiency, including its ability to work efficiently on mobile devices.

Compression Rate. Throughout our experiments, we define the compression rate as the ratio of compressed data size to its original counterpart, with units based on file size. In our model performance comparisons (Sec. V-D1), the compression rate refers to the reduction applied to Mel spectrograms—for instance, a 12.5% rate means that the compressed feature has 12.5% the size of the original Mel spectrogram. In contrast, for the system overhead evaluation (Sec. V-D2), the compression rate is calculated relative to the raw audio waveform file size. Under this interpretation, a 12.5% Mel spectrogram compression corresponds approximately to 6.25% of the original waveform size, as the Mel spectrogram typically accounts for around half the storage size of the full audio signal in our pipeline.

Accuracy. Accuracy is the fundamental metric for evaluating classification tasks, including speech emotion recognition. It measures the proportion of correct predictions (both true positives and true negatives) out of all predictions made. Specifically, accuracy is defined as:

$$Accuracy = \frac{Number\,of\,Correct\,Predictions}{Total\,Number\,of\,Predictions} \tag{9}$$

In the context of speech emotion recognition, accuracy quantifies how well the model identifies the correct emotional label for the input speech samples. A high accuracy score indicates that the model can effectively classify emotional states from speech data.

F1 Score. While accuracy is important, it may not be sufficient in scenarios with imbalanced data or when certain

emotional classes dominate. To address this, we utilize the F1-score as a more balanced metric, which considers both precision and recall. The F1-score is the harmonic mean of precision and recall and is defined as:

$$F1 = \frac{2 \cdot (precision \cdot recall)}{precision + recall} \tag{10}$$

The precision represents the fraction of relevant instances among the retrieved instances, while recall measures the fraction of relevant instances that were retrieved. The F1-score provides a better understanding of the model's ability to handle class imbalances by rewarding models that balance both precision and recall. This ensures that the model doesn't simply predict the majority class with high accuracy, but also correctly identifies minority classes, which is crucial for emotion recognition tasks.

Energy Consumption In addition to classification performance, we evaluate energy consumption during audio stream transmission, which is crucial for mobile devices. The energy consumption is measured during the transfer of compressed audio features from the mobile device to the server. This metric accounts for both the computational load and network overhead involved in transmitting the audio stream. Lower energy consumption ensures that the model can run efficiently on mobile devices without draining the battery, which is especially important for real-time emotion recognition tasks.

Latency We also measure the latency associated with audio stream transmission, which refers to the time taken to process the input audio signal and transmit it for emotion recognition. Latency is critical in real-time applications like speech emotion recognition, where quick feedback is required. Our system aims to minimize latency by efficiently transmitting compressed audio features and processing them on the server. Low latency ensures that predictions are made promptly, providing a smooth user experience.

D. Micro Benchmark

1) Model Comparison: The experiments in this section validate the emotion recognition accuracy by comparing different deep neural networks, including UMONS [67], Xu [26], and Yoon [27], [66]. Additionally, to investigate the impact of signal compression on speech emotion recognition, we evaluate the recognition accuracy under different compression rates.

First, comparing the proposed system with other networks, we observe that incorporating external world knowledge into our deep neural network significantly improves emotion recognition performance. The accuracy achieved by our system is 73.92% on IEMOCAP and 67.40% on VCEMO for the 4-way classification, which is notably higher than the accuracy of the baseline networks, demonstrating the effectiveness of integrating additional contextual knowledge. This result confirms that the proposed model can better capture the complex nuances of emotional expression in speech, leading to superior performance.

Moreover, we analyze the effect of compression rate on the system's recognition accuracy. As the compression rate increases, the accuracy only shows a slight decrease from 73.92% to 72.85% on IEMOCAP and 67.40% to 65.23 % on VCEMO, which is still significantly higher than the accuracy of the other networks across all compression rates (see Tab. I and Tab. III). This indicates that our system is robust to compression, maintaining high performance even when speech features are heavily compressed. Such resilience to compression is crucial for mobile and edge devices, where bandwidth and storage are limited.

The system's performance is further validated by the F1-score, which is a combination of accuracy and recall, providing a more comprehensive evaluation of the model's effectiveness. As shown in Tab. II and Tab. IV, the F1-scores of our model consistently outperform the existing emotion recognition methods, further solidifying the advantage of our approach in terms of both precision and recall. This demonstrates that the proposed system not only achieves high accuracy but also performs well in terms of balancing false positives and false negatives, which is essential for real-world emotion recognition tasks.

In summary, the experimental results show that our system Amser+, while leveraging signal compression and external knowledge, achieves superior emotion recognition accuracy and remains robust under varying compression rates. This makes our approach highly suitable for resource-constrained environments, such as mobile devices and edge computing systems, without compromising recognition performance.

TABLE I
ACCURACY COMPARISON OF DIFFERENT MODELS ON IEMOCAP

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**72.85%**	67.84%	63.34%	55.52%
18.75%	**73.25%**	67.64%	63.64%	55.91%
25.00%	**73.52%**	67.64%	63.74%	56.21%
50.00%	**73.81%**	67.74%	63.83%	56.89%
100.00%	**73.92%**	67.64%	64.32%	58.26%

TABLE II
F1 SCORE COMPARISON OF DIFFERENT MODELS ON IEMOCAP

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**0.713**	0.677	0.630	0.548
18.75%	**0.716**	0.675	0.633	0.553
25.00%	**0.721**	0.676	0.633	0.556
50.00%	**0.725**	0.677	0.635	0.564
100.00%	**0.728**	0.675	0.640	0.577

TABLE III
ACCURACY COMPARISON OF DIFFERENT MODELS ON VCEMO

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**65.23%**	61.26%	58.12%	59.82%
18.75%	**65.87%**	61.54%	58.43%	59.79%
25.00%	**66.92%**	62.57%	58.67%	60.42%
50.00%	**66.85%**	63.06%	59.03%	60.85%
100.00%	**67.40%**	63.27%	59.42%	60.96%

2) System Overhead: In this section, we investigate the effect of signal compression on power consumption and latency during the transmission of compressed speech data over WiFi. The experiment utilizes a compression rate of 6.25%, with a total of 22,366 audio files being transferred for evaluation. The latency and energy consumption of different compression methods are measured and compared, including

TABLE IV
F1 SCORE COMPARISON OF DIFFERENT MODELS ON VCEMO

Compress Rate	Ours	UMONS	Xu	Yoon
12.50%	**0.654**	0.621	0.542	0.572
18.75%	**0.664**	0.624	0.563	0.575
25.00%	**0.668**	0.628	0.572	0.584
50.00%	**0.670**	0.631	0.571	0.583
100.00%	**0.672**	0.634	0.577	0.591

TABLE V
TRANSMISSION TIME AND ENERGY CONSUMPTION

System	RAW	MP3	AAC	Intp	Amser+
Latency(s)	406.58	357.82	344.15	276.38	180.75
Energy(kWh)	0.0056	0.0051	0.0044	0.0041	0.0025

RAW, MP3, AAC, and Interpolation (Intp), along with our proposed system, Amser+.

As shown in Tab. V, the transmission latency and energy consumption of each system are summarized. When transferring raw audio files (RAW), the latency is 406.58 seconds, and the energy consumption is 0.0056 kWh. In contrast, the systems using lossy compression methods such as MP3 and AAC achieve faster transmission times, with latencies of 357.82 seconds and 344.15 seconds, respectively, and slightly reduced energy consumption compared to RAW. The Interpolation-based compression (Intp) method further reduces latency to 276.38 seconds, with energy consumption decreasing to 0.0041 kWh.

However, our proposed system, Amser+, demonstrates the most significant improvements. It reduces latency to 180.75 seconds, which is 2.24 times faster than RAW transmission and 1.53 times faster than the interpolation-based method. Furthermore, our system achieves a 55.35% reduction in energy consumption, dropping to just 0.0025 kWh compared to the raw transmission, marking a substantial improvement in both latency and energy efficiency.

These results highlight the effectiveness of our proposed system in optimizing both the speed and energy efficiency of speech data transmission. The combination of signal compression and advanced feature extraction techniques enables our system to achieve faster processing times while significantly lowering the power consumption, making it highly suitable for resource-constrained environments, such as mobile and edge devices, where both energy efficiency and latency are critical factors.

E. Ablation study

To further understand the effect of each modality, we performed an ablation study based on the 6.25% compression rate. The results are presented in Tab. VI and Tab. VII, where we test the network model's performance in emotion recognition using different modalities.

When using only Mel Features, the system achieves 55.18% accuracy and an F1-score of 0.541 on IEMOCAP, and 51.85% / 0.503 on VCEMO. This shows that Mel features, which capture acoustic properties like pitch and tone, provide some useful emotional cues, but they are not sufficient on their own for optimal emotion recognition.

TABLE VI
ABLATION STUDY OF USING DIFFERENT MODALITIES ON IEMOCAP: EMBEDDINGS MEANS THE SIMPLE TRANSCRIPTION EMBEDDINGS WHILE THE ROBERTA MEANS THE ROBERTA EMBEDDINGS.

Used modality	**Accuracy**	**F1-score**
Embeddings	59.40%	0.571
RoBERTa	55.29%	0.532
Mel Features	55.18%	0.541
Embeddings + RoBERTa	59.26%	0.581
Embeddings + Mel Features	68.21%	0.674
RoBERTa + Mel Features	64.23%	0.639
Embeddings + RoBERTa + Mel Features	**72.85%**	**0.713**

TABLE VII
ABLATION STUDY OF USING DIFFERENT MODALITIES ON VCEMO: EMBEDDINGS MEANS THE SIMPLE TRANSCRIPTION EMBEDDINGS WHILE THE ROBERTA MEANS THE ROBERTA EMBEDDINGS.

Used modality	**Accuracy**	**F1-score**
Embeddings	55.82%	0.541
RoBERTa	51.96%	0.498
Mel Features	51.85%	0.503
Embeddings + RoBERTa	56.01%	0.539
Embeddings + Mel Features	64.13%	0.632
RoBERTa + Mel Features	59.89%	0.596
Embeddings + RoBERTa + Mel Features	**65.23%**	**0.654**

a) Impact of Word Embedding.: Adding Word Embeddings leads to significant improvements. The combination of Mel Features and Word Embeddings achieves 68.21% accuracy (F1: 0.674) on IEMOCAP and 64.13% (F1: 0.632) on VCEMO. This suggests that lexical features extracted from transcriptions effectively complement acoustic signals across both English and Mandarin speech.

b) Impact of RoBERTa.: When Mel Features and RoBERTa are combined, the accuracy increases to 64.23% on IEMOCAP and 59.89% on VCEMO. This result indicates that while RoBERTa's contextual knowledge enhances the feature extraction from speech, it works better when fused with Mel Features than when used alone. However, it still underperforms compared to the combination of Mel Features + Word Embeddings, suggesting that the latter provides more direct and relevant features for emotion recognition.

Finally, the best performance on both datasets is achieved by integrating all three modalities. On IEMOCAP, the full fusion yields 72.85% accuracy and an F1-score of 0.713; on VCEMO, the system reaches 65.23% and 0.654, respectively. These results underscore the importance of multimodal fusion in capturing diverse emotional signals across languages and speaking styles.

VI. USER STUDY

In this section, we examine the usability of Amser+. We invited 10 participants (7 male, 3 female, ages 20–35) to use the Amser+ that performed real-time speech emotion recognition and provided feedback in the form of emotion labels. After completing a short task (e.g., reading predefined utterances and observing system responses), participants were asked to rate the system using the 12-item System Usability Scale (SUS [78]) questionnaire on a 5-point Likert scale ranging from "strongly agree" to "strongly disagree". The questionnaire is as follows:

About you

In this section, you will be presented with a number of questions about yourself.

1 Your gender:
- ☐ Female
- ☐ Male

2 How old are you?
- ☐ Below 18
- ☐ 18-24
- ☐ 25-34
- ☐ Above 35

3 Your education level
- ☐ High school or lower
- ☐ Undergraduate degree
- ☐ Master's degree
- ☐ Doctoral degree
- ☐ Other: _____

4 Languages you speak fluently:
- ☐ English
- ☐ Chinese
- ☐ Hindi
- ☐ Spanish
- ☐ French
- ☐ Other: _____

About the user experience

In this section, you will be presented with a number of questions about your opinions and attitudes towards Amser+.

After completing the short task using the proposed *Amser+* system, please rate your opinions on the following aspects.

1 I think I would like to use the Amser+ emotion recognition system frequently in real-world applications.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

2 I found the emotion recognition process more complicated than necessary.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

3 I thought the system was easy to interact with using my voice.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

4 I think I would need help from a technical person to use this system.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

5 I felt that processing and feedback were well integrated in this system.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

6 I noticed inconsistencies in the way the system responded to different emotional tones.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

7 I believe most users would learn how to use the system quickly.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

8 I found the overall interaction with the system to be cumbersome.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

9 I felt confident that the emotional labels predicted by the system matched my intended emotion.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

10 I had to figure out too many things before I could use the system properly.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

11 I feel the system responded quickly enough to be used in real-time interactions.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

12 I noticed a noticeable delay between my speech and the system's feedback.
Strongly Disagree (-2) ☐—☐—☐—☐—☐ Strongly Agree (2)

Please describe your experience with Amser+. What aspects did you particularly enjoy? Are there any areas where you think the system could be improved?

Table VIII summarizes participants' responses to the 12-item usability questionnaire. The results demonstrate an overall positive perception of the Amser+ system.

For positively worded items (Q1, Q3, Q5, Q7, Q9, Q11), participants reported high levels of agreement. In particular, Q7 and Q3 received the highest average ratings of 1.1 and 1.0, respectively, indicating the system's learnability and interaction friendliness. Q11, which specifically evaluated real-time responsiveness, also scored highly (1.0), suggesting that most users found the system sufficiently fast for real-time applications.

For negatively worded items (Q2, Q4, Q6, Q8, Q10, Q12), the average scores were consistently below zero; after reverse scoring, this indicates favorable user sentiment. For instance, Q10 received the lowest raw average (-1.2), suggesting that participants strongly disagreed with the statement and found the system easy to start using. Similarly, Q12 scored -0.9, indicating that most users did not perceive any significant latency during interaction.

Taken together, these results indicate that Amser+ demonstrates strong usability, with both direct and reverse-scored items showing consistent user satisfaction.

VII. Discussion

A. Noise Robustness in Real-World Scenarios.

In practical scenarios such as in-car environments, open offices, or smart homes, ambient noise is often inevitable and

TABLE VIII
THE RESULT OF QUESTIONNAIRE.

	Strongly Disagree (-2)	Disagree (-1)	Not sure (0)	Agree (1)	Strongly Agree (2)	Average Rating
Q1	1	1	2	2	4	0.7
Q2	4	2	2	2	0	-0.8
Q3	0	1	3	1	5	1.0
Q4	3	3	3	1	0	-0.8
Q5	1	2	1	2	4	0.6
Q6	4	1	1	3	1	-0.4
Q7	0	1	1	4	4	1.1
Q8	5	2	2	1	0	-1.1
Q9	2	1	1	4	2	0.3
Q10	5	2	3	0	0	-1.2
Q11	0	1	1	5	3	1.0
Q12	4	2	3	1	0	-0.9

varies significantly in intensity and spectral characteristics. To mitigate the impact of noise, our system employs Mel-spectrogram features, which are perceptually motivated and emphasize frequency components most relevant to human auditory perception. This naturally suppresses irrelevant high-frequency noise and improves robustness in moderately noisy conditions.

Moreover, in real-world applications where noise levels are particularly high, Amser+ can be extended with a lightweight real-time speech enhancement module on the mobile device such as [31], [32], [79]. This module, running in parallel with audio acquisition, would enhance human speech signals and attenuate background noise before feature extraction. Such integration is orthogonal to our current compression and recognition pipeline and can further improve performance without modifying the downstream network. Exploring such enhancements remains an important direction for future work in extreme acoustic environments.

B. Handling Overlapping Speech.

The Amser+ system is not explicitly designed to address overlapping speech scenarios, where multiple speakers speak simultaneously within the same audio segment. Such situations are common in real-world mobile contexts, such as cafés, group conversations, or shared office spaces. In the presence of overlapping speech, the input signal may contain mixed acoustic and emotional cues from different speakers, which poses significant challenges for both automatic speech recognition (ASR) and emotion classification. ASR performance may degrade due to misalignment of speaker-dependent phonetic content, and emotion recognition may fail to isolate speaker-specific emotional expressions, resulting in ambiguous or inaccurate predictions. Effectively handling overlapping speech would require incorporating techniques such as speech separation (e.g., source separation networks) or speaker diarization, which can isolate individual speaker streams from a mixture. These components, however, introduce additional model complexity and may require speaker-level supervision or computational resources not yet optimized for mobile deployment. We consider this an important but orthogonal extension to our current system design and leave it as a promising direction for future work aiming at enhanced multi-speaker robustness.

C. Personalization and Speaker Adaptation

In mobile environments where frequent users interact with the system, personalization can play a crucial role in improving recognition accuracy and user satisfaction. Although Amser+ is designed as a speaker-independent model to ensure broad generalizability, its modular structure allows for the potential integration of user-specific adaptation. For instance, lightweight fine-tuning or embedding adaptation techniques could be applied to the ASR or emotion recognition modules based on a user's historical speech data. Such personalization has been shown to improve recognition accuracy by better capturing individual speech traits, emotional expression styles, and linguistic patterns. Moreover, recent advances in on-device continual learning and speaker embedding-based personalization provide promising pathways for incremental, privacy-preserving adaptation without compromising latency or model efficiency. However, we note that there is currently a lack of large-scale, long-duration datasets from single users that would enable comprehensive personalization studies. We leave the collection and evaluation of such data as an important direction for future work to enhance system robustness and personalization in long-term deployments.

VIII. CONCLUSION

We propose Amser+, a real-time speech emotion recognition framework for mobile devices. By offloading deep neural network computations to a server, the system reduces the load on mobile devices. Speech signals are compressed using Fbank features and SVD, minimizing storage requirements while preserving key emotional cues. A pre-trained RoBERTa model further enhances emotion recognition accuracy by incorporating external knowledge. Extensive experiments validate the system's feasibility, showing it achieves high accuracy and low energy consumption, making it ideal for mobile-based emotion recognition applications.

ACKNOWLEDGMENTS

This work is supported in part by National Natural Science Foundation of China (No. 61936015), Natural Science Foundation of Shanghai (No. 24ZR1430600) and Shanghai Key Laboratory of Trusted Data Circulation and Governance, and Web3.

REFERENCES

[1] fortunebusinessinsights, "speech-and-voice-recognition-market-101382," https://www.fortunebusinessinsights.com/industry-reports/speech-and-voice-recognition-market-101382, 2023.

[2] A. B. Ingale and D. Chaudhari, "Speech emotion recognition," *International Journal of Soft Computing and Engineering (IJSCE)*, vol. 2, no. 1, pp. 235–238, 2012.

[3] K. Wu and D. G. Childers, "Gender recognition from speech. part i: Coarse analysis," *The journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1828–1840, 1991.

[4] D. G. Childers and K. Wu, "Gender recognition from speech. part ii: Fine analysis," *The Journal of the Acoustical society of America*, vol. 90, no. 4, pp. 1841–1856, 1991.

[5] M. Zhao, F. Adib, and D. Katabi, "Emotion recognition using wireless signals," in *Proceedings of the 22nd annual international conference on mobile computing and networking*, 2016, pp. 95–108.

[6] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, and I. Sutskever, "Robust speech recognition via large-scale weak supervision," in *International Conference on Machine Learning*. PMLR, 2023, pp. 28 492–28 518.

[7] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[8] N. Goyal, J. Du, M. Ott, G. Anantharaman, and A. Conneau, "Larger-scale transformers for multilingual masked language modeling," *arXiv preprint arXiv:2105.00572*, 2021.

[9] W. Zhang, Z. He, L. Liu, Z. Jia, Y. Liu, M. Gruteser, D. Raychaudhuri, and Y. Zhang, "Elf: accelerate high-resolution mobile deep vision with content-aware parallel offloading," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 201–214.

[10] Y. Lu, D. Ding, H. Pan, Y. Fu, L. Zhang, F. Tan, R. Wang, Y.-C. Chen, G. Xue, and J. Ren, "M3cam: Extreme super-resolution via multi-modal optical flow for mobile cameras," in *Proceedings of the 22nd ACM Conference on Embedded Networked Sensor Systems*, 2024, pp. 744–756.

[11] Y. Lu, H. Pan, F. Tan, Y.-C. Chen, J. Yu, J. He, and G. Xue, "Effectively learning moiré qr code decryption from simulated data," in *IEEE INFOCOM 2023-IEEE Conference on Computer Communications*. IEEE, 2023, pp. 1–10.

[12] Z. Liu, M. Sun, T. Zhou, G. Huang, and T. Darrell, "Rethinking the value of network pruning," *arXiv preprint arXiv:1810.05270*, 2018.

[13] K. Ullrich, E. Meeds, and M. Welling, "Soft weight-sharing for neural network compression," *arXiv preprint arXiv:1702.04008*, 2017.

[14] K. Dokic, M. Martinovic, and D. Mandusic, "Inference speed and quantisation of neural networks with tensorflow lite for microcontrollers framework," in *2020 5th South-East Europe Design Automation, Computer Engineering, Computer Networks and Social Media Conference (SEEDA-CECNSM)*. IEEE, 2020, pp. 1–6.

[15] J. H. Cho and B. Hariharan, "On the efficacy of knowledge distillation," in *Proceedings of the IEEE/CVF international conference on computer vision*, 2019, pp. 4794–4802.

[16] D. Ding, L. Yang, Y.-C. Chen, and G. Xue, "Leakage or identification: Behavior-irrelevant user identification leveraging leakage current on laptops," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 5, no. 4, Dec. 2022. [Online]. Available: https://doi.org/10.1145/3494984

[17] W. Park, D. Kim, Y. Lu, and M. Cho, "Relational knowledge distillation," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 3967–3976.

[18] J. Michalakes and M. Vachharajani, "Gpu acceleration of numerical weather prediction," in *2008 IEEE International Symposium on Parallel and Distributed Processing*. IEEE, 2008, pp. 1–7.

[19] S. Li, C. Wu, H. Li, B. Li, Y. Wang, and Q. Qiu, "Fpga acceleration of recurrent neural network based language model," in *2015 IEEE 23rd Annual International Symposium on Field-Programmable Custom Computing Machines*. IEEE, 2015, pp. 111–118.

[20] E. Nurvitadhi, J. Sim, D. Sheffield, A. Mishra, S. Krishnan, and D. Marr, "Accelerating recurrent neural networks in analytics servers: Comparison of fpga, cpu, gpu, and asic," in *2016 26th International Conference on Field Programmable Logic and Applications (FPL)*. IEEE, 2016, pp. 1–4.

[21] D. Wu, A. Chen, T. E. Ng, G. Wang, and H. Wang, "Accelerated service chaining on a single switch asic," in *Proceedings of the 18th ACM Workshop on Hot Topics in Networks*, 2019, pp. 141–149.

[22] K. Baker, "Singular value decomposition tutorial," *The Ohio State University*, vol. 24, p. 511, 2005.

[23] S. Ghosh, U. Tyagi, S. Ramaneswaran, H. Srivastava, and D. Manocha, "Mmer: Multimodal multi-task learning for speech emotion recognition," *arXiv preprint arXiv:2203.16794*, 2022.

[24] I. Gat, H. Aronowitz, W. Zhu, E. Morais, and R. Hoory, "Speaker normalization for self-supervised speech emotion recognition," in *ICASSP 2022*. IEEE, 2022, pp. 7342–7346.

[25] A. Triantafyllopoulos, S. Liu, and B. W. Schuller, "Deep speaker conditioning for speech emotion recognition," in *ICME*. IEEE, 2021, pp. 1–6.

[26] H. Xu, H. Zhang, K. Han, Y. Wang, Y. Peng, and X. Li, "Learning alignment for multimodal emotion recognition from speech," *Proc. Interspeech 2019*, pp. 3569–3573, 2019.

[27] S. Yoon, S. Byun, S. Dey, and K. Jung, "Speech emotion recognition using multi-hop attention mechanism," in *ICASSP 2019*. IEEE, 2019, pp. 2822–2826.

[28] Y. Wang, A. Boumadane, and A. Heba, "A fine-tuned wav2vec 2.0/hubert benchmark for speech emotion recognition, speaker verification and spoken language understanding," *arXiv preprint arXiv:2111.02735*, 2021.

[29] C. Busso, M. Bulut, C.-C. Lee, A. Kazemzadeh, E. Mower, S. Kim, J. N. Chang, S. Lee, and S. S. Narayanan, "Iemocap: Interactive emotional dyadic motion capture database," *Language resources and evaluation*, vol. 42, pp. 335–359, 2008.

[30] K. He, H. Fan, Y. Wu, S. Xie, and R. Girshick, "Momentum contrast for unsupervised visual representation learning," in *CVPR*, 2020, pp. 9729–9738.

[31] A. Pandey and D. Wang, "Tcnn: Temporal convolutional neural network for real-time speech enhancement in the time domain," in *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2019, pp. 6875–6879.

[32] K. Tan and D. Wang, "A convolutional recurrent neural network for real-time speech enhancement." in *Interspeech*, vol. 2018, 2018, pp. 3229–3233.

[33] Y. Lu, R. Wang, D. Ding, H. Zhang, L. Zhang, L. Yang, Y.-C. Chen, and G. Xue, "Amser: Accelerate mobile speech emotion recognition with signal compression," in *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. IEEE, 2025, pp. 1–5.

[34] H. Pan, L. Qiu, B. Ouyang, S. Zheng, Y. Zhang, Y.-C. Chen, and G. Xue, "Pmsat: Optimizing passive metasurface for low earth orbit satellite communication," in *Proceedings of the 29th Annual International Conference on Mobile Computing and Networking*, 2023, pp. 1–15.

[35] Y. Song, H. Pan, L. Ge, L. Qiu, S. Kumar, and Y.-C. Chen, "Microsurf: Guiding energy distribution inside microwave oven with metasurfaces," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 1346–1360.

[36] R. Ma, S. Zheng, H. Pan, L. Qiu, X. Chen, L. Liu, Y. Liu, W. Hu, and J. Ren, "Automs: Automated service for mmwave coverage optimization using low-cost metasurfaces," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 62–76.

[37] Y. Fu, S. Wang, L. Zhong, L. Chen, J. Ren, and Y. Zhang, "Ultrasr: Silent speech reconstruction via acoustic sensing," *IEEE Transactions on Mobile Computing*, vol. 23, no. 12, pp. 12 848–12 865, 2024.

[38] ——, "Svoice: Enabling voice communication in silence via acoustic sensing on commodity devices," in *Proceedings of the 20th ACM Conference on Embedded Networked Sensor Systems*, ser. SenSys '22. New York, NY, USA: Association for Computing Machinery, 2023, p. 622–636. [Online]. Available: https://doi.org/10.1145/3560905.3568530

[39] Y. Fu, Y. Zhang, H. Pan, Y. Lu, X. Li, L. Chen, J. Ren, X. Li, X. Zhang, and Y. Zhang, "Pushing the limits of acoustic spatial perception via incident angle encoding," *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, vol. 8, no. 2, May 2024. [Online]. Available: https://doi.org/10.1145/3659583

[40] Y. Fu, Y. Zhang, Y. Lu, L. Qiu, Y.-C. Chen, Y. Wang, M. Wang, Y. Li, J. Ren, and Y. Zhang, "Adaptive metasurface-based acoustic imaging using joint optimization," in *Proceedings of the 22nd Annual International Conference on Mobile Systems, Applications and Services*, ser. MOBISYS '24. New York, NY, USA: Association for Computing Machinery, 2024, p. 492–504. [Online]. Available: https://doi.org/10.1145/3643832.3661863

[41] H. Pan, Y. Wang, J. Liu, R. Ma, L. Qiu, Y.-C. Chen, G. Xue, and J. Ren, "Cgmm: Non-invasive continuous glucose monitoring in wearables using metasurfaces," in *Proceedings of the 31th Annual International Conference on Mobile Computing and Networking*, 2025, pp. 1–16.

[42] H. Kong, L. Lu, J. Yu, Y. Chen, X. Xu, and F. Lyu, "Toward multi-user authentication using wifi signals," *IEEE/ACM Transactions on Networking*, vol. 31, no. 5, pp. 2117–2132, 2023.

[43] M. Chen, Y. Ma, Y. Li, D. Wu, Y. Zhang, and C.-H. Youn, "Wearable 2.0: Enabling human-cloud integration in next generation healthcare systems," *IEEE Communications Magazine*, vol. 55, no. 1, pp. 54–61, 2017.

[44] H. Kong, C. Huang, J. Yu, and X. Shen, "A survey of mmwave radar-based sensing in autonomous vehicles, smart homes and industry," *IEEE Communications Surveys & Tutorials*, 2024.

[45] L. Fridman, P. Langhans, J. Lee, and B. Reimer, "Driver gaze region estimation without use of eye movement," *IEEE Intelligent Systems*, vol. 31, no. 3, pp. 49–56, 2016.

[46] K. Meesublak and T. Klinsukont, "A cyber-physical system approach for predictive maintenance," in *2020 ieee international conference on smart internet of things (smartiot)*. IEEE, 2020, pp. 337–341.

[47] Y. Wang, H. Pan, L. Qiu, L. Zhong, J. Liu, R. Ma, Y.-C. Chen, G. Xue, and J. Ren, "Gpms: Enabling indoor gnss positioning using passive metasurfaces," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 1424–1438.

[48] M. M. Baig, H. GholamHosseini, and M. J. Connolly, "Mobile healthcare applications: system design review, critical issues and challenges," *Australasian physical & engineering sciences in medicine*, vol. 38, pp. 23–38, 2015.

[49] Y. Wang and Y.-C. Chen, "Non-contact thermal haptics for vr," in *Adjunct Proceedings of the 2023 ACM International Joint Conference on Pervasive and Ubiquitous Computing & the 2023 ACM International Symposium on Wearable Computing*, 2023, pp. 386–390.

[50] R. T. Azuma, "A survey of augmented reality," *Presence: teleoperators & virtual environments*, vol. 6, no. 4, pp. 355–385, 1997.

[51] L. Chen, B. Yu, Y. Fu, J. Ren, H. Pan, J. Gummeson, and Y. Zhang, "Pushing wireless charging from station to travel," in *Proceedings of the 30th Annual International Conference on Mobile Computing and Networking*, 2024, pp. 46–61.

[52] H. Pan and L. Qiu, "Passive metasurface-based low earth orbit ground station design," *Tsinghua Science and Technology*, vol. 30, no. 1, pp. 148–160, 2024.

[53] ——, "Passive metasurface for interacting with electromagnetic signals," Sep. 19 2024, uS Patent App. 18/608,421.

[54] T. Zhang, A. Chowdhery, P. Bahl, K. Jamieson, and S. Banerjee, "The design and implementation of a wireless video surveillance system," in *Proceedings of the 21st Annual International Conference on Mobile Computing and Networking*, 2015, pp. 426–438.

[55] Y. Li, A. Padmanabhan, P. Zhao, Y. Wang, G. H. Xu, and R. Netravali, "Reducto: On-camera filtering for resource-efficient real-time video analytics," in *Proceedings of the Annual conference of the ACM Special Interest Group on Data Communication on the applications, technologies, architectures, and protocols for computer communication*, 2020, pp. 359–376.

[56] C. Canel, T. Kim, G. Zhou, C. Li, H. Lim, D. G. Andersen, M. Kaminsky, and S. Dulloor, "Scaling video analytics on constrained edge nodes," *Proceedings of Machine Learning and Systems*, vol. 1, pp. 406–417, 2019.

[57] S. Naderiparizi, P. Zhang, M. Philipose, B. Priyantha, J. Liu, and D. Ganesan, "Glimpse: A programmable early-discard camera architecture for continuous mobile vision," in *Proceedings of the 15th Annual International Conference on Mobile Systems, Applications, and Services*, 2017, pp. 292–305.

[58] J. Emmons, S. Fouladi, G. Ananthanarayanan, S. Venkataraman, S. Savarese, and K. Winstein, "Cracking open the dnn black-box: Video analytics with dnns across the camera-cloud boundary," in *Proceedings of the 2019 workshop on hot topics in video analytics and intelligent edges*, 2019, pp. 27–32.

[59] S. Jiang, Z. Lin, Y. Li, Y. Shu, and Y. Liu, "Flexible high-resolution object detection on edge devices with tunable latency," in *Proceedings of the 27th Annual International Conference on Mobile Computing and Networking*, 2021, pp. 559–572.

[60] B. Schuller, G. Rigoll, and M. Lang, "Hidden markov model-based speech emotion recognition," in *ICASSP*, vol. 2. IEEE, 2003, pp. II–1.

[61] G. Trigeorgis, F. Ringeval, R. Brueckner, E. Marchi, M. A. Nicolaou, B. Schuller, and S. Zafeiriou, "Adieu features? end-to-end speech emotion recognition using a deep convolutional recurrent network," in *ICASSP*. IEEE, 2016, pp. 5200–5204.

[62] T. Seehapoch and S. Wongthanavasu, "Speech emotion recognition using support vector machines," in *2013 KST*. IEEE, 2013, pp. 86–91.

[63] C.-C. Lee, E. Mower, C. Busso, S. Lee, and S. Narayanan, "Emotion recognition using a hierarchical binary decision tree approach," *Speech Communication*, vol. 53, no. 9-10, pp. 1162–1171, 2011.

[64] M. M. El Ayadi, M. S. Kamel, and F. Karray, "Speech emotion recognition using gaussian mixture vector autoregressive models," in *ICASSP*, vol. 4. IEEE, 2007, pp. IV–957.

[65] A. Stuhlsatz, C. Meyer, F. Eyben, T. Zielke, G. Meier, and B. Schuller, "Deep neural networks for acoustic emotion recognition: Raising the benchmarks," in *ICASSP*. IEEE, 2011, pp. 5688–5691.

[66] S. Yoon, S. Byun, and K. Jung, "Multimodal speech emotion recognition using audio and text," in *2018 SLT*. IEEE, 2018, pp. 112–118.

[67] J.-B. Delbrouck, N. Tits, M. Brousmiche, and S. Dupont, "A transformer-based joint-encoding for emotion recognition and sentiment analysis," *ACL 2020*, p. 1, 2020.

[68] A. Dax, "The eckart-young theorem and ky fan's maximum principle: Two sides of the same coin," in *Householder Symposium XVIII on Numerical Linear Algebra*. Citeseer, 2011, p. 49.

[69] V. Panayotov, G. Chen, D. Povey, and S. Khudanpur, "Librispeech: an asr corpus based on public domain audio books," in *2015 IEEE international conference on acoustics, speech and signal processing (ICASSP)*. IEEE, 2015, pp. 5206–5210.

[70] R. Ardila, M. Branson, K. Davis, M. Henretty, M. Kohler, J. Meyer, R. Morais, L. Saunders, F. M. Tyers, and G. Weber, "Common voice: A massively-multilingual speech corpus," *arXiv preprint arXiv:1912.06670*, 2019.

[71] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation," in *EMNLP*, 2014, pp. 1532–1543.

[72] Z. Yu, J. Yu, Y. Cui, D. Tao, and Q. Tian, "Deep modular co-attention networks for visual question answering," in *CVPR*, 2019, pp. 6281–6290.

[73] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, vol. 30, 2017.

[74] P. Khosla, P. Teterwak, C. Wang, A. Sarna, Y. Tian, P. Isola, A. Maschinot, C. Liu, and D. Krishnan, "Supervised contrastive learning," *NeurIPS*, vol. 33, pp. 18 661–18 673, 2020.

[75] J. Tang, L. Zhang, Y. Lu, D. Ding, L. Yang, Y. Chen, M. Bian, X. Li, and G. Xue, "Vcemo: Multi-modal emotion recognition for chinese voiceprints," *arXiv preprint arXiv:2408.13019*, 2024.

[76] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.

[77] A. Paszke, S. Gross, F. Massa *et al.*, "Pytorch: An imperative style, high-performance deep learning library," *Advances in neural information processing systems*, vol. 32, 2019.

[78] S. U. S. (SUS). [Online]. Available: https://www.usability.gov/how-to-andtools/methods/system-usability-scale.html.

[79] Y. R. Pei, R. Shrivastava, and F. Sidharth, "Real-time speech enhancement on raw signals with deep state-space modeling," *arXiv preprint arXiv:2409.03377*, 2024.